Fast Inverse Square Root $(0x5F3759df)$

floating point $f_x = \underbrace{\boxed{S_x \mid E_x \mid M_x}}^{N\text{-bits}}$ — S_x: 1 bit, E_x: b bits, M_x: $N-b-1$ bits

Sqrt only defined for positive values so $S_x = 0$

$$\therefore f_x = \left(1 + \frac{M_x}{2^{N-b-1}}\right) 2^{E_x - (2^{b-1}-1)}$$

For simplicity, let $L = 2^{N-b-1}$, which normalizes the mantissa M to $0 \leq M < 1$

& let $B = 2^{b-1} - 1$, which is the exponent bias.

$$f_x = \left(1 + \frac{M_x}{L}\right) 2^{E_x - B} \qquad (1)$$

Looking for $y = \frac{1}{\sqrt{x}} = x^{-\frac{1}{2}}$

Let f_x & f_y be floating point representations of x & y respectively.

$\therefore f_y = f_x^{-\frac{1}{2}}$ ignoring errors introduced by floating pt. approx.

$$\log_2 f_y = \log_2 \left(f_x^{-\frac{1}{2}}\right)$$

$$\log_2 f_y = -\tfrac{1}{2} \log_2 f_x$$

$$\log_2\left(\left(1 + \frac{M_y}{L}\right) 2^{E_y - B}\right) = -\tfrac{1}{2} \log_2\left(\left(1 + \frac{M_x}{L}\right) 2^{E_x - B}\right)$$

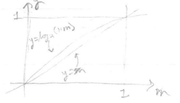
$$\log_2\left(1 + \frac{M_y}{L}\right) + \log_2\left(2^{E_y - B}\right) = -\tfrac{1}{2}\left[\log_2\left(1 + \frac{M_x}{L}\right) + \log_2\left(2^{E_x - B}\right)\right]$$

$$\log_2\left(1 + \frac{M_y}{L}\right) + E_y - B = -\tfrac{1}{2} \log_2\left(1 + \frac{M_x}{L}\right) - \tfrac{1}{2} E_x + \tfrac{1}{2} B$$

$$\log_2\left(1+\frac{M_y}{L}\right)+E_y=-\frac{1}{2}\log_2\left(1+\frac{M_x}{L}\right)-\frac{1}{2}E_x+\frac{3}{2}B$$

$$\log_2\left(1+\frac{M_y}{L}\right)+E_y=-\frac{1}{2}\log_2\left(1+\frac{M_x}{L}\right)-\frac{1}{2}(E_x-3B) \qquad (2)$$

Now consider the binary log: $\log_2(1+m)$ for $0 \le m < 1$:



We see that $\log_2(1+m) \approx m$ for $0 \le m < 1$

More specifically, $\log_2(1+m) = m + \theta_m$ for $0 \le m < 1$ and some small error term θ_m (3)

$\therefore$ Substituting (3) into (2) we have:

$$\frac{M_y}{L}+\theta_y+E_y=-\frac{1}{2}\left(\frac{M_x}{L}+\theta_x\right)-\frac{1}{2}(E_x-3B)$$

$$\frac{M_y}{L}+E_y=-\frac{1}{2}\frac{M_x}{L}-\frac{1}{2}\theta_x-\theta_y-\frac{1}{2}(E_x-3B)$$

$$E_yL+M_y=-\frac{1}{2}M_x-L\left(\frac{1}{2}\theta_x+\theta_y\right)-\frac{1}{2}L(E_x-3B)$$

$$=-\frac{1}{2}M_x-L\left(\frac{1}{2}\theta_x+\theta_y\right)-\frac{1}{2}E_xL+\frac{3}{2}BL$$

$$E_yL+M_y=L\left(\frac{3}{2}B-\left(\frac{1}{2}\theta_x+\theta_y\right)\right)-\frac{1}{2}(E_xL+M_x) \qquad (4)$$

Now lets take a quick look at an n-bit floating point register organization:

Cast to an integer type, this is $E2^{N-b-1}+M=EL+M$

So if we take I_x and I_y to be the floating point values f_x & f_y, respectively, cast to integers, then we have

$$I_x = E_x L + M_x \quad , \quad I_y = E_y L + M_y. \tag{5}$$

So from (5), (4) turns into

$$I_y = L\left(\tfrac{3}{2}B - \left(\tfrac{1}{2}\vartheta_x + \vartheta_y\right)\right) - \tfrac{1}{2} I_x$$

or more simply

$$I_y = R - \tfrac{1}{2} I_x$$

where $R = L\left(\tfrac{3}{2}B - \left(\tfrac{1}{2}\vartheta_x + \vartheta_y\right)\right)$

And that's the basic technique: ~~take~~ floating point x as an integer, divide it by 2 (right shift by 1), & subtract it from our magic integer R. This gives I_y, & you take that as a floating point you get y.

The error comes from the fact that ϑ_x & ϑ_y are for specific values of x (& therefore of y).

So we need to pick ϑ_x & ϑ_y as best as we can, & it will be an approximation for most values of x.

For single precision IEEE floating point, $N=32$, $e=8$

$L = 2^{23}$

$B = 127$

$$R = 2^{23}\left(\tfrac{3}{2}(127) - \left(\tfrac{1}{2}\vartheta_x + \vartheta_y\right)\right)$$